Title: Fast Track: fast, (nearly) automatic formant-tracking using Praat

Abstract

Fast Track is a formant tracker implemented in Praat that attempts to automatically select the best analysis from a set of candidates. The best track is selected by modeling smooth formant contours across the entirety of the sound, providing the researcher with rich information about static and dynamic formant properties. Fast Track returns text files containing acoustic information (e.g, formants, formant bandwidths, fundamental frequency) sampled every 2 ms, generates images showing the winning analysis and comparing alternate analyses, and creates log files detailing analysis information for each file. Fast Track features a modular workflow that allows for analysis steps to be run (and re-run) independently as necessary. In addition, Fast Track is designed to allow for easy correction of tracking errors by allowing the user to override the automatic analysis, or manually edit tracks where necessary. The design and use of Fast Track are outlined using a re-analysis of the Hillenbrand et al. (1995) dataset, which suggests that Fast Track can be very accurate in cases where signal properties allow for reliable formant estimates.

1. Introduction

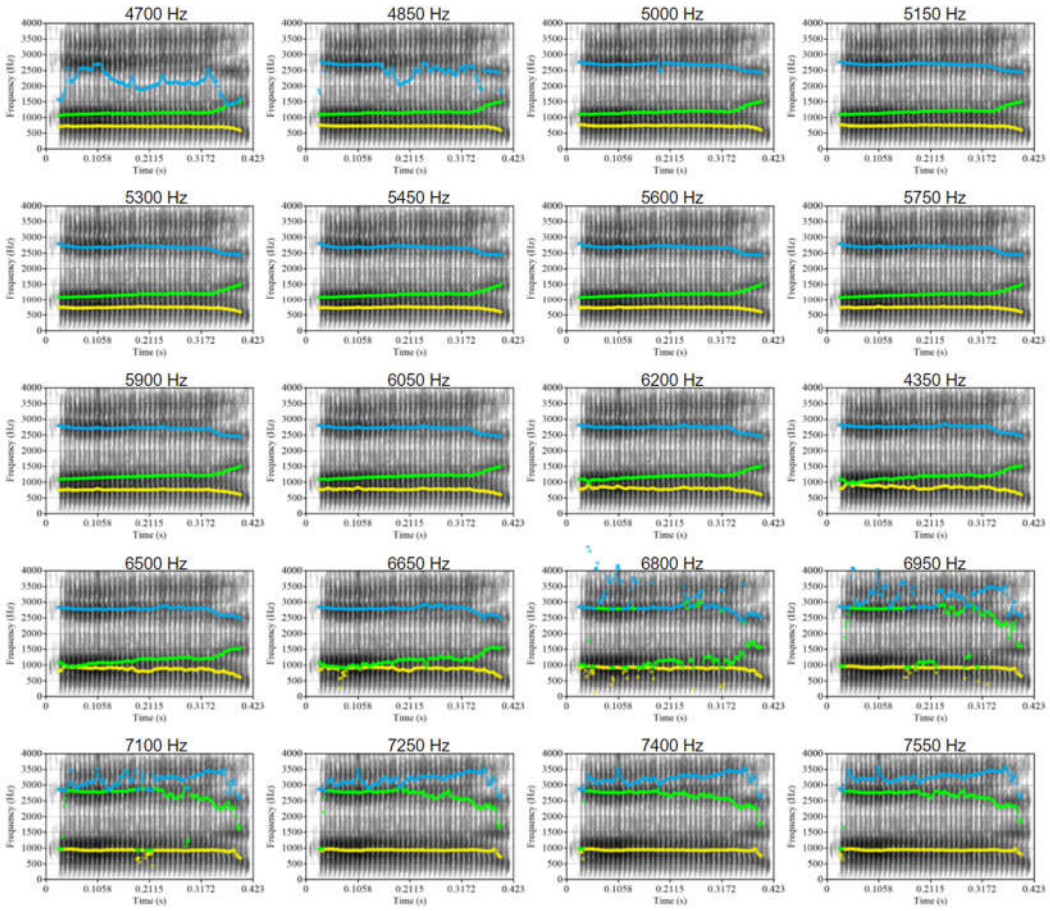
Linguists often measure the formant frequencies of vowel sounds in order to make quantitative statements about dialectal and social differences between speakers. Although reliable inferences depend on accurate formant measurements, formant frequencies are notoriously difficult to measure automatically with high accuracy. The main obstacle to accurate, unsupervised formant-estimation is that the optimal analysis parameters vary between speakers and phones (Kendall & Vaughn, 2020). As a result, it is difficult to know a priori which analysis best suits a given token. In the simplest approach, the researcher can repeatedly adjust the analysis settings (i.e., attempt multiple analyses) until they find an acceptable result. However, this approach is time consuming, difficult to reproduce, and not systematic. In response to this, many automatic formant-tracking schemes have been proposed over the years (e.g., Nearey et al., 2002; Escudero et al., 2009; Zhang et al., 2013; Weenink, 2015). Most of these follow the same general procedure: 1) A set of candidate analyses are generated by modifying analysis parameters in a systematic way, 2) The 'goodness/badness' of each analysis is quantified using one or more metrics. 3) The 'best' analysis is selected based on the metric(s) and returned to the user. Effectively, these approaches automate the repeated cycles of parameter adjustments and output evaluation that would otherwise be carried out by the researcher.

Fast Track is a plugin for Praat (Boersma, 2020) that provides a modular, customizable, automatic formant-tracker. Fast Track carries out multiple analyses for each sound (as in Figure 1) and looks for the 'best' analysis. Fast Track will then generate text files containing time-varying acoustic information (formant frequencies, formant bandwidths, fundamental frequency, sound intensity, and harmonicity), and can be modified to include any other analysis carried out by Praat. This rich information allows researchers to model formant contours, and to summarize formant values with precise control. Fast Track is implemented entirely using Praat scripts, which are saved as plain-text files. This means that the user can easily change or extend its behavior by modifying the appropriate script in any text-editing software. In addition, although Fast Track can be quite accurate in many cases, a focus has been placed on creating an easy workflow in situations where automatic tracking fails. Fast Track makes correct analyses easy to verify, and errors easy to correct by saving information regarding all intermediate analysis steps, and all alternate analyses.

2. Fast Track

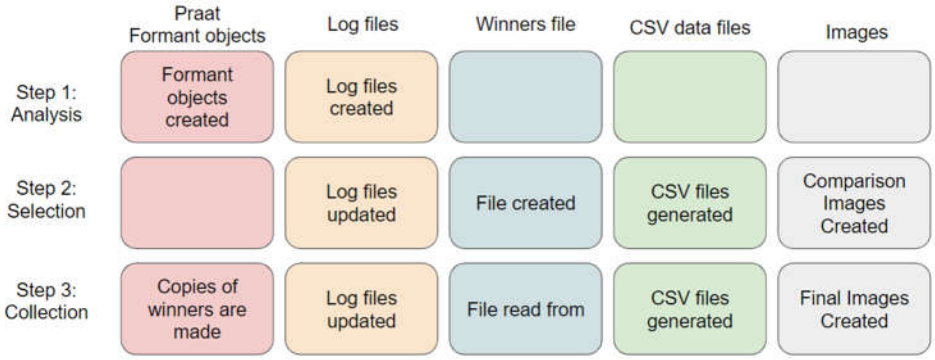
An overview of the design and function of Fast Track will be provided here. For more specific information about the current implementation, parameter settings, or general usage, please visit the Fast Track wiki [link]. Fast Track is intended to analyze sound files that contain only a single vowel sound[1], and can analyze a single file at a time or carry out a batch analysis of an entire folder of files (each containing a single sound). The description of the folder-tracking option will be provided in this section, with a summary of the data generated at each step provided in Figure 2.

[1] More generally, the method used to pick the best analysis assumes that formants exist and are continuous throughout the recording. Fast Track can be used to analyze sounds that contain sequences of sonorants and vowels, or potentially other sounds, provided that these feature formants without discontinuities, and that the order of the regression analysis is large enough to adequately represent the movement of the formants in the interval.



63 *Figure 1 - Twenty analyses of the same sound: a production of 'had' by an adult male speaker. Each analysis looks for the same number of formants (5.5), though only the lowest three are drawn. Analyses vary in the maximum frequency below which the algorithm looks for formants (indicated above each plot), increasing in 150 Hz steps*
66 *from left to right, top to bottom. A broad range of intermediate frequencies provide accurate analyses. Cutoff frequencies that are too low (top left, 4700 Hz) or too high (bottom right, 7550 Hz) can result in large errors.*

69



72 *Figure 2 - Schematic of the output files created and modified at each step of Fast Track. In the first step, analysis files are generated, in the second step they are evaluated, and in the final step the best analyses are collected. More information is provided in sections 2.1, 2.2, and 2.3.*

2.1 Analyzing the sounds

Linear Predictive Coding (LPC) is used to estimate formant frequencies. An LPC analysis assumes that the sound spectrum can be characterized by a series of formants, and then tries to estimate the frequencies (and bandwidths) of those formants. A researcher using LPC has to set two parameters: How many formants to look for and what frequency to look below? Following Praat, these parameters will be referred to as 'Maximum number of formants' and 'Maximum formant (Hz)', respectively. If a researcher looks for more formants than exist below a given maximum-formant frequency, LPC will tend to find formants where there are none. For example, much of the F3 track in the top-left plot in Figure 1 is in a location between F2 and F3 where there is no formant. When the researcher asks for fewer formants than exist below a given frequency, LPC will tend to skip or merge formants, as has happened with F1 and F2 in the bottom-right of Figure 1. Thus, accurate LPC analyses require the user to specify an appropriate number of formants to be found below a given maximum-formant frequency.

In the first step, the candidate formant-analyses are generated and recorded for each sound file. Fast track requires that audio files be cut so that they each contain only a single vowel, and provides functions to help extract vowels from longer stretches of audio. Although Fast Track allows the user to collect either 3 or 4 formants, the initial analysis always looks for 5.5 formants (the half formant helps model variation in spectral tilt). Modeling 5 formants even if the researcher desires fewer ensures that the formants that are of interest will be substantially lower than the maximum-formant frequency. Fast Track asks the user to specify low and high maximum-formant frequencies, and the number of steps that should be taken between these limits.

For instance, a user might indicate that they want to take 20 analysis steps between maximum-formant frequencies of 4800 and 7300 Hz (including the endpoints). This means that Fast Track will analyze a token looking for 5.5 formants below 4800 Hz, and save the analysis in a text file representing a Praat 'Formant' object. The maximum-formant frequency will then be increased to the next step (in this case 4931 Hz) and the analysis will again be carried out and recorded. This process will be repeated until the final analysis is reached (maximum formant = 7300 Hz). The output of this stage is *n* text files for each sound file, representing the *n* analysis steps. Fast Track also generates log files containing analysis information for each sound file.

2.2. Selecting the best analysis

Fast Track selects the best analysis based primarily on the intrinsic properties of the formant tracks, with only some (optional) heuristics to rule out unrealistic values (e.g., median F1 > 1200 Hz). The goodness of different analyses is quantified using a similar approach to Weenink (2016): each formant is modeled using a regression analysis, and the smoothness of each track is quantified using the residuals of the regressions (see Figure 3). Since samples are taken densely across time (every 2 ms by default), and each sound file only contains a single vowel segment, variation from sample to sample that is not along a smooth trajectory is very likely to indicate a measurement error. Thus, all other things being equal, the analysis with the smallest residuals (i.e. the least 'roughness') across all

formants will reflect the smoothest formant trajectories and will be selected as the best track[2]. In cases where an analysis violates one of the heuristic constraints, a penalty is added to the roughness score, further decreasing the viability of a given analysis. For analyses of equal smoothness, the one that violates the fewest heuristic constraints will be chosen.

In the second step, the following actions are taken for each sound file. Regression models are fit for each formant, predicting frequency as a function of time (predicted formant values and regression coefficients are recorded in log files). After this, the total 'badness' score (roughness + heuristic violations) for each analysis frequency is found and recorded in a text file. A comma-separated file (a file containing the 'winners') is generated recording the best analysis for each sound (i.e., the smallest badness score), allowing this information to be viewed and modified by the user. Finally, images are generated comparing all of the candidate analyses and indicating which analysis was automatically selected as best (similar to Figure 1).

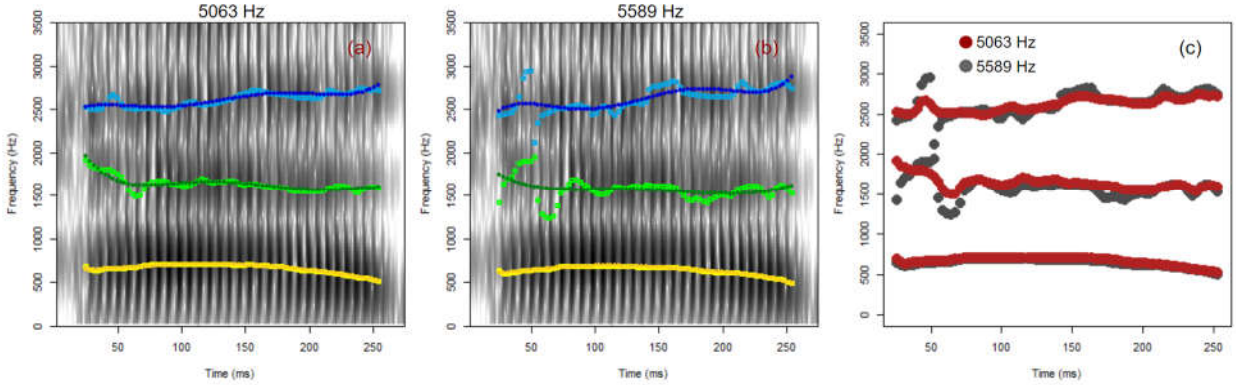


Figure 3 - (a) Result of analyzing an adult-male production of 'head' with a maximum formant of 5063 Hz. Filled points indicate predicted formant values, empty points (in lighter colors) indicate observed formant frequencies. The mean absolute error (MAE) between predicted and observed values is 58 Hz. (b) An analysis of the same sound with a maximum formant of 5589 Hz, resulting in an MAE of 151 Hz. (c) A comparison of the two analyses, 5063 Hz in red, and 5589 Hz in grey. Despite their similarities, the analysis in red is smoother, and therefore preferable.

2.3 Collecting winners and generating output data

In the third step, Fast Track refers to the text file recording the best analysis for each sound file (the 'winners' file generated at the end of the second step). For example, if the winners file indicates that the best analysis for the first sound is analysis 14, Fast Track will make a copy of this analysis file and move it to a directory that contains *only* the winning analysis for each file. The user also has the option of specifying a different analysis for each formant, which Fast Track will combine into a single analysis. For example, a researcher can replace a single bad formant in an otherwise good analysis by combining the F1, and F2 from one analysis with the F3 from another analysis. After this, Fast Track will generate a comma-separated text file containing the following information sampled every 2 ms (by default): formant frequencies and bandwidths, predicted formant frequencies,

[2] Since Fast Track selects the single best overall analysis across all formants, this may not provide the best analysis individually for each specific formant. As discussed in Section 2.3, Fast Track allows the user to combine formant tracks from different maximum-formant frequencies. Fast Track also offers several ways to model the formant trajectories and to calculate and combine smoothness estimate, as discussed in the wiki [link]).

fundamental frequency, sound intensity, and harmonicity. An image is also generated comparing a spectrogram of each sound to the observed and predicted formant frequencies suggested by the winning analysis.

Although the second step (selecting winners) and third step (collecting winners) of the analysis are run automatically by default, Fast Track allows these steps to be run (and re-run) separately. This separation facilitates the correction of errors without requiring any technical knowledge, and allows for work to easily be carried out in a distributed fashion. For example, researchers can validate analyses by looking at the generated images, and these decisions can be recorded into the text file containing information about winning analyses with any text-editing software. Alternatively, researchers may wish to select winners entirely by other means, and then integrate these selections into their Fast Track workflow. In addition to the intrinsic properties of different analyses, the selection of the best analysis could also incorporate knowledge of plausible formant ranges for phonemes in a language or dialect (e.g., Labov et al., 2013). For example, since an F1 of 700 Hz is not plausible for /i/, any analysis returning such a value can be ruled out as a candidate. When enough tokens are available for each speaker, the selection of the best analysis could also consider expectations regarding values for each vowel and formant (e.g., Escudero et al., 2009). For instance, if a speaker tends to produce F3 near 2400 Hz for /o/, a single F3 near 3400 Hz is likely to be an error. Although Fast Track does not incorporate such methods into its initial selection, information regarding the formant contours and smoothness of all alternate analyses is saved in text files. As a result, the analyses generated by Fast Track can easily be imported into statistical computing environments such as R, Python or Excel, where more complex selection methods can be implemented. Once the user has determined a final set of best analyses and saved these to the text file recording this information, the third step can be run (or re-run) so that Fast Track will generate the final data files, log files, and images.

2.4 Editing analyses and summarizing data

Fast Track has several tools to help manipulate or summarize data[3]. In some cases, a winning analysis can still benefit from hand-editing, and thus it is particularly useful to be able to modify analyses efficiently. Fast Track takes advantage of Praat's built-in functionalities to allow users to manually edit tracks quickly and easily by converting data into editable FormantGrid objects (see Figure 4), and then generating data files, images and Formant objects from the edited data. Fast Track keeps copies of both the edited and original Formant objects so that the unedited data can be recovered later.

Although information about formant contours can be very useful, formant estimates every 2 ms can be 'too much' information for some purposes. Fast Track has several functions to help summarize data across tokens. For example, the mean (or median) values of formant frequencies within equally-sized subsections of each vowel can be calculated and collected across tokens. Such data summaries can help produce vowel plots that present the same data in different ways, as in Figure 5, without ever requiring that the data be re-analyzed.

[3] For a current outline of available tools see the Fast Track wiki [link].

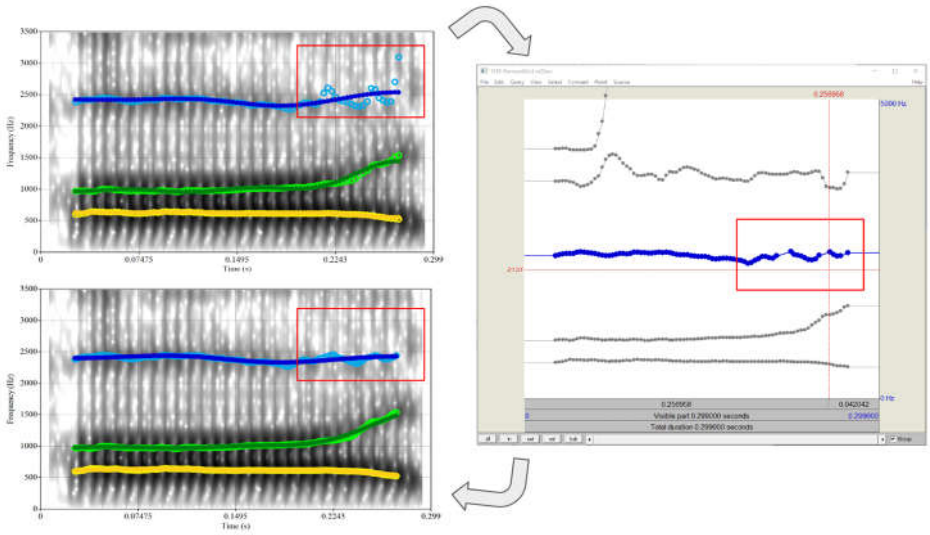


Figure 4 - The top spectrogram presents an analysis with some noisy measurements towards the end of F3. On the right, the track has been converted to a FormantGrid object so that the points that deviate substantially from the predicted values can be deleted. The end result (bottom spectrogram) is a final analysis with more reliable F3 values.

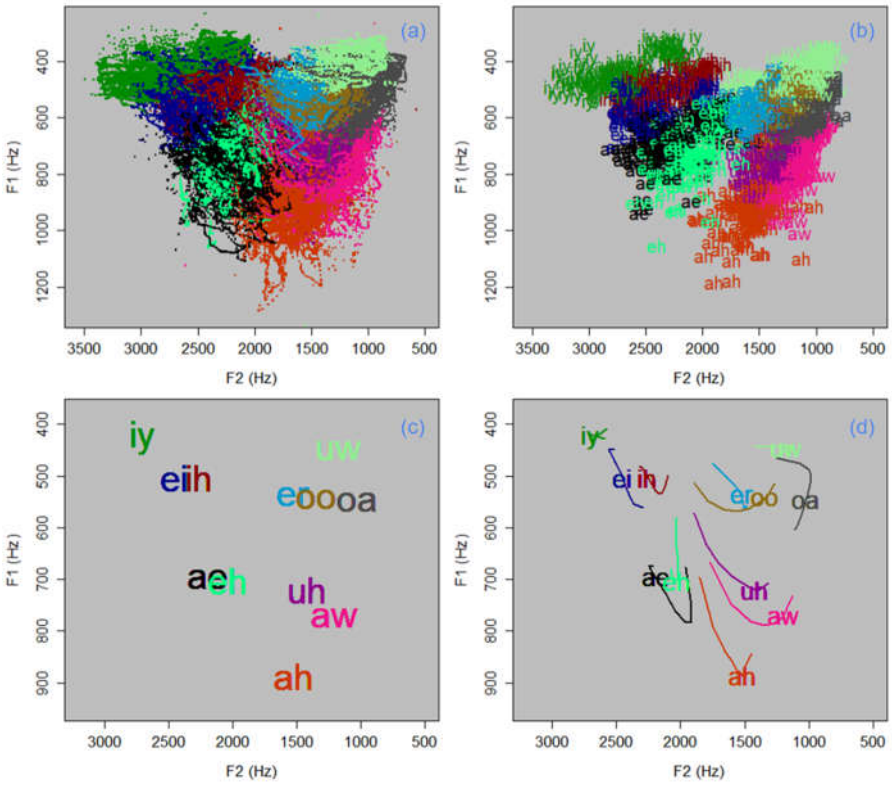


Figure 5 - (a) All individual observations (n ≈ 150,000) of F1 and F2 across all tokens in the Hillenbrand et al. (1995) data for the re-analysis described in Section 3. (b) Average F1 and F2 values from midpoints (40-50% of vowel durations), for each token (n = 1668). (c) Average midpoint values for each vowel category, across all tokens. (d) Curves indicate F1 and F2 values for each vowel, averaged across every tenth of vowel duration, averaged across all speakers. Labels have been placed at midpoints.

3. Performance evaluation

Performance will be evaluated by carrying out a re-analysis of the Hillenbrand et al (1995) dataset (henceforth H95, available here). The H95 data consists of a single repetition of 12 vowels (/i, ɪ, ʊ, u, e, o, ɛ, ʌ, ɔ, æ, ɑ, ɝ/) in an /hVd/ context by 139 speakers: 48 adult females, 45 adult males, 19 girls and 27 boys (all 10-12 years old). Vowel portions of the sound files were extracted using the beginning and end times reported by the authors. To investigate the relationship between maximum-formant settings and analysis error for different speakers, a 'gold-standard' set of formant tracks is required to represent the ideal analysis in the opinion of the user. Selection of the gold-standard analysis was done by generating a set of analyses varying in maximum-formant, for each token. Twenty analyses were considered, with maximum-formant frequency varying between 4800 and 7300 Hz, resulting in an analysis every 131 Hz (i.e., maximum formant = 4800, 4931, 5062, …, 7300 Hz). Images comparing the 20 analyses were made for each sound in a similar fashion to Figure 1, and the best analysis (in the opinion of the author) was selected as the gold-standard.

3.1 Maximum-formant settings for different speaker types

It may be tempting to carry out an analysis with an overly broad maximum-formant range in order to ensure that the correct analysis is included in the set of candidates. However, as seen in Figure 6, maximum-formant settings that are very inappropriate for a given sound can result in analyses that are extremely inaccurate, and yet very difficult to distinguish from a good formant-track. In many cases, this is because the incorrect analysis resembles a production of a different phone by another type of speaker. For example, the formant pattern suggested by the analysis in Figure 6b could plausibly be a low back-vowel produced by an adult (it is reasonably similar to the true formant pattern in Figure 6c), and is smoother than the correct analysis in Figure 6a. Often in situations like these, the only way to select the correct analysis is to refer to information external to the token being analyzed (e.g., the formants in 6b are not plausible for an open vowel produced by a child).

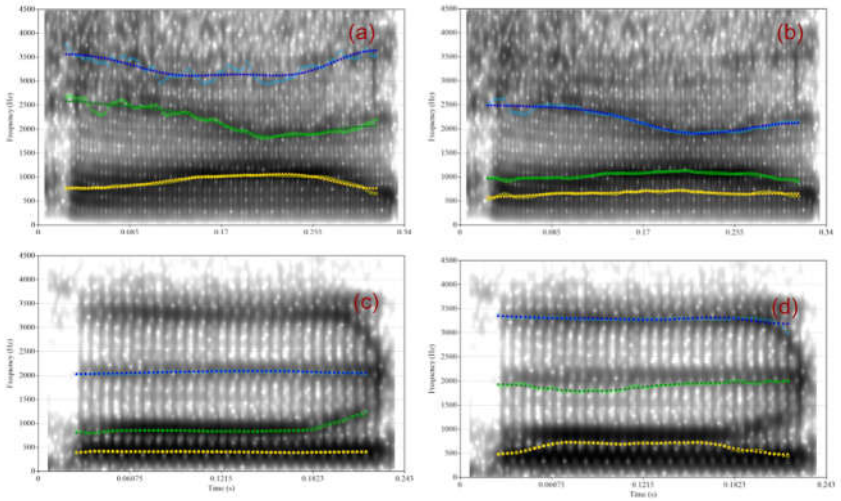


Figure 6 - (a) The gold-standard analysis of a production of 'had' by a girl (maximum formant = 6653 Hz). (b) The same sound as (a), but the maximum-formant is set much too low (4700 Hz). (c) The gold-standard analysis of a production of 'who'd' by an adult male (maximum formant = 5147 Hz). (d) The same sound as (c), but the maximum-formant is set much too high (6547 Hz). Filled points indicate predicted formant values, empty points (in lighter colors) indicate observed formant frequencies

To investigate the relationship between maximum-formant frequency settings and analysis error, the following calculation was carried out for each speaker. First, for each token, the mean absolute error (MAE) was found between the gold-standard values for each formant at each time point, and the corresponding values for the analysis at each maximum-formant frequency setting. Then, these values were averaged within-speaker (across vowels), resulting in a single value for each speaker for each maximum-formant setting (20 values per speaker). These MAEs answer the following question for a given speaker: How much error can we expect in our analysis for each maximum-formant setting? Plots showing variation in MAE as a function of maximum-formant frequency (Figure 7) can be used to understand which analysis ranges best suit different sorts of speakers. For example, in Figure 7d we see that there are large measurement errors for maximum-formant settings below 5500 Hz for most girls in the H95 data, suggesting that maximum-formant frequencies below 5500 Hz should likely be avoided for these speakers. Although there are obvious general differences between speaker classes, there is also substantial variability within-class. This suggests that upper and lower bounds on maximum-formant frequencies should be set based on the characteristics of the particular speaker rather than solely due to belonging to a general class (e.g., 'woman', see Figure 7f).

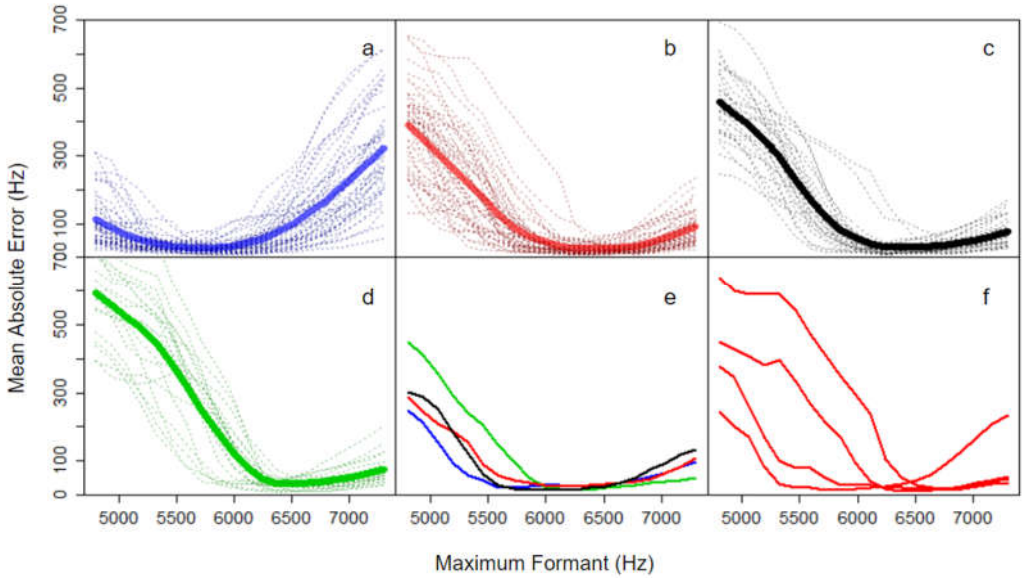


Figure 7 - Bold lines indicates average for speaker class, dotted lines indicate average errors for individual speakers for (a) adult males, (b) adult females, (c) 10-12 year-old boys, (d) 10-12 year-old girls. (e) Curves comparing one speaker from each class, showing these can be quite similar across gross speaker classes. (f) Curves comparing four adult females show that within-class variation can be large.

3.1 Expected analysis error under different selection conditions

In addition to selecting the maximum-formant range for the analyses, the user has control over the selection of the winner from among the candidates. To investigate the reliability of analyses in different selection conditions, the following possibilities will be considered:

1. Twenty analyses from 4800 to 7300 Hz are considered. The smoothest track is selected as the winner (smoothness only).
2. The same analyses from (1) are considered. In addition to smoothness, the following heuristics are used: Median F1 should be under 1200 Hz, and the median bandwidths of F1, F2 and F3 should be under 500, 600, and 900 Hz, respectively. Finally, if median F3 is under 2000 Hz, then the difference between median F2 and median F1 should be greater than 500 Hz. This final heuristic allows extremely low F3 values only when F1 and F2 feature enough separation.
3. Different maximum-formant ranges: Twenty analyses between 4800 and 6600 Hz are used for adult males, twenty analyses between 5500 Hz and 7300 Hz are used if the speaker is not an adult male. The smoothest track is selected as the winner.
4. In addition to smoothness, the heuristics from (2) are combined with the different analysis ranges from (3).
5. Fixed frequencies are used for tokens produced by all speakers in a given class based on the default Praat values: 5000 Hz for adult males and 5500 Hz for all other speakers.

Rather than looking for an exact match, the number of cases where the automatic analysis is reasonably similar to the gold-standard will be measured. Similarity was quantified by calculating two statistics: 1) the mean absolute-error (MAE) between the gold standard and the automatically-selected winner for each token, 2) the percent of tokens where the MAE is less than 5% of the value of the gold-standard formant frequencies (since measurement error tends to increase proportionally with formant frequency, Hillenbrand et al. 1995, Table II). Results are presented in Figure 8.

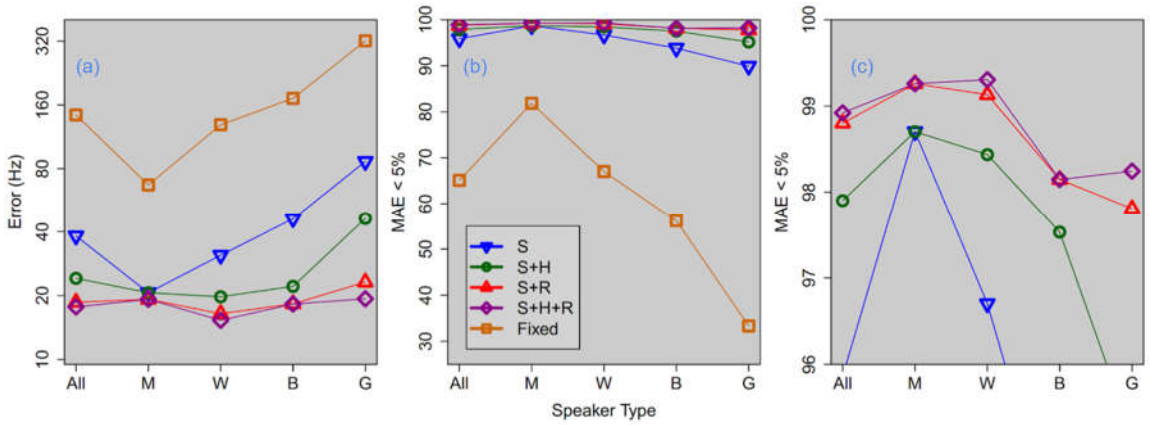


Figure 8 - (a) Mean absolute error (MAE) between winners and gold-standards for different speaker classes (Men, Women, Boys, Girls), for different selection methods. Selection methods combined considerations of track smoothness (S), heuristics (H), and range differences (R) between adult male and non-adult male speakers. Note the y-axis has logarithmic spacing. (b) Percent of analyses where the MAE is less than 5% of the gold-standard values. (c) The same information as in (b), but with a smaller y-axis range.

Overall, the combination of range restrictions and heuristics resulted in an average error of around 20 Hz, and 1650/1668 tokens having an error smaller than 5% of values. Simply providing appropriate ranges for speakers (method 3 above, S+R in Figure 8) resulted in a substantial

performance improvement over only considering formant smoothness (method 1, S in Figure 8), highlighting the importance of providing appropriate maximum-formant ranges for an analysis. The static analysis frequencies resulted in very poor performance, while combinations of the heuristics and range restrictions yielded very good results. It should be noted that performance could have been better, perhaps much better, with more appropriate static frequencies. The problem is it is difficult to know what these should be a priori, thus leading to the desire for automation in the analysis.

4. Summary and conclusion

The general design and use of Fast Track was outlined here. The modular design allows for the distribution of work in a laboratory, and the use of intermediate data files means that Fast Track can easily integrate with alternative analysis methods. Fast Track makes it simple to collect large amounts of data automatically (formant frequencies, formant bandwidths, f0, sound intensity, and harmonicity measured every 2 ms), and can be modified to include any other analysis carried out by Praat. Because it is implemented entirely in Praat, researchers need no special technical ability to use it, while those comfortable with Praat scripting can easily modify and extend its behavior.

References

Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.16, retrieved 6 June 2020 from [http://www.praat.org/]

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393. https://doi.org/10.1121/1.3180321

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Kendall, T., & Vaughn, C. (2020). Exploring vowel formant estimation through simulation-based techniques. *Linguistics Vanguard*, *6*(s1). https://doi.org/10.1515/lingvan-2018-0060

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language*, *89*(1), 30–65. https://doi.org/10.1353/lan.2013.0015

Nearey, T. M., Assmann, P. F., & Hillenbrand, J. M. (2002). Evaluation of a strategy for automatic formant tracking. *The Journal of the Acoustical Society of America*, *112*(5), 2323–2323. https://doi.org/10.1121/1.4779372

Weenink, D., & others. (2015). Improved formant frequency measurements of short segments. *ICPhS*.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication*, *55*(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011